

大学・研究機関連携で作る新しい 学術情報基盤mdxとそのネットワーク構成

ADVNET 2021

東京大学 情報基盤センター 助教

中村 遼

データ活用社会創成プラットフォーム

「データ活用社会創成プラットフォーム」は用途に応じてオンデマンドで**短時間に構築・拡張・融合**できるデータ収集・集積・解析機能を提供するプラットフォーム。

研究所 (2)

・ 国立情報学研究所、産業技術総合研究所

大学 (9)

・ 北海道大、東北大、筑波大、東京大、東京工業大、名古屋大、京都大、大阪大、九州大



データ活用社会創成プラットフォーム 3本柱

1

SINETを活かしたリアルタイム収集・集積・解析環境の動的な構築

遠隔地のセンサーやストレージ、データプラットフォームの計算資源、ストレージをつないで、リアルタイムに入力から出力を得られるアプリケーションごとの収集・集積・解析環境（仮想データプラットフォーム：仮想DP）を、使いたいときに即時に構築する
SINETモバイル基盤によりセンサー等のデータを安定してセキュアにつなぐ

2

高性能計算環境によるデータ科学と計算科学の融合

データ科学、計算科学の手法を融合し、さらに国内最高の計算環境を用いて他に無い高精度の予測を行えるようにする

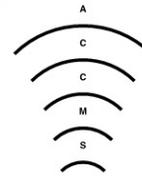
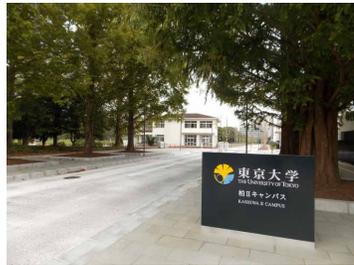
3

異種データ・異種知識の融合活用の推進と利用者支援

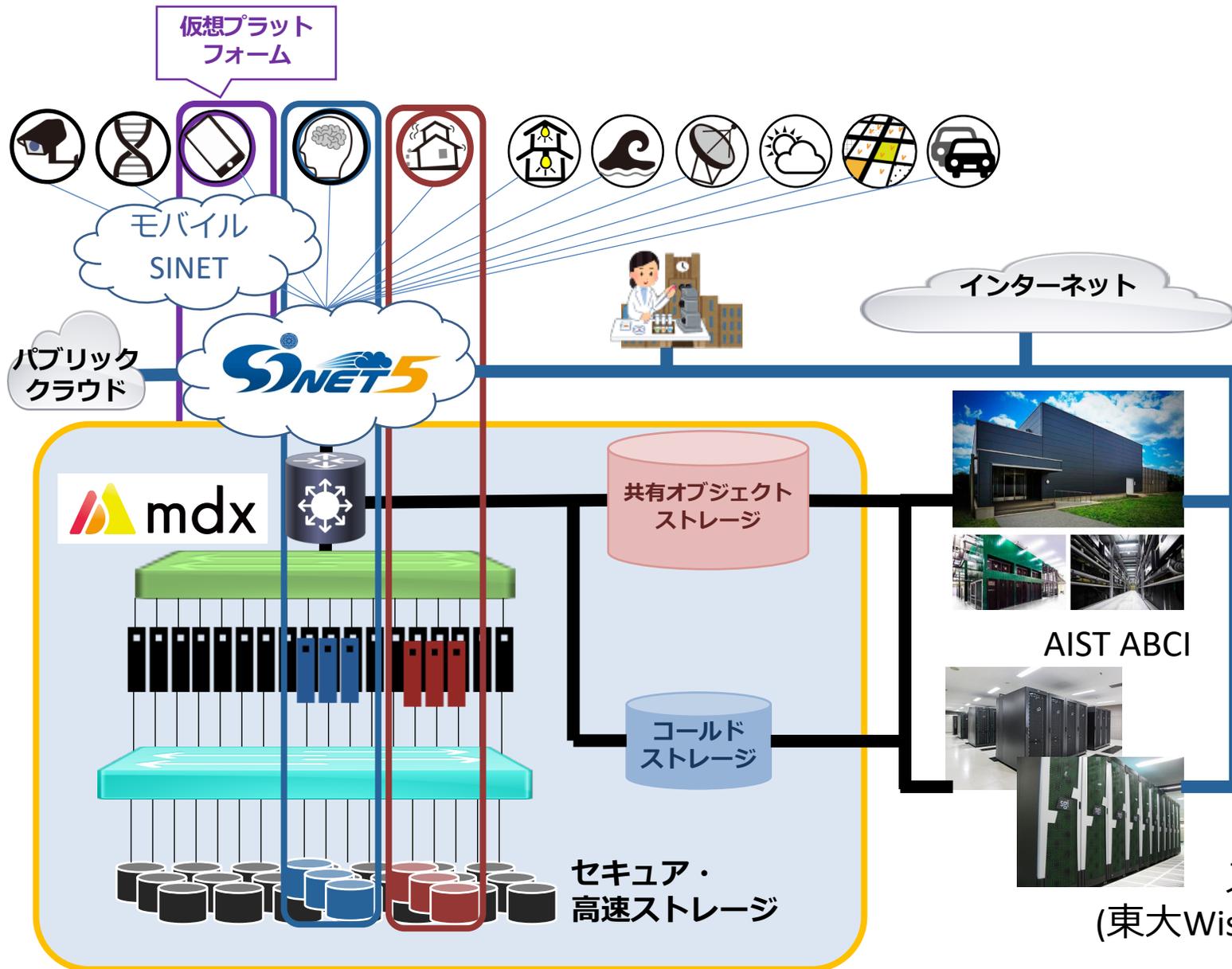
様々な分野のデータ保持者、解析者、利用者が産学にまたがって連携するコミュニティを形成し、新たな価値創造につなげる。
データ活用を目指す利用者へのコンサルティングや開発支援を実施する。

mdxとは

- 9大学2研究所が共同運営し、全国共同利用に供する、データ活用にフォーカスした高性能IaaS環境
- 東京大学 柏2キャンパスに設置



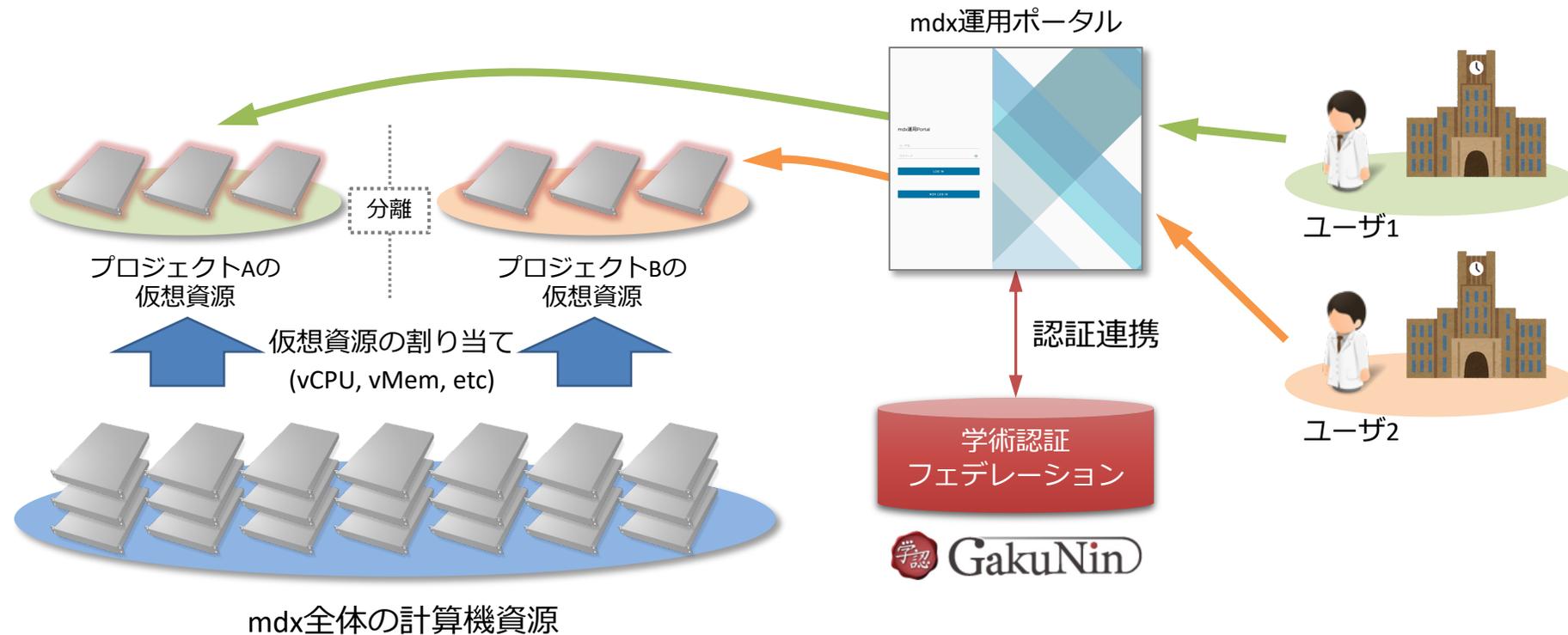
mdxの概要



- **仮想プラットフォーム**
 - 柔軟・セキュアな環境の構築が可能
 - 高性能な「マルチテナント」環境
- **大容量ストレージ**
 - セキュア
 - データ共有
- **SINET・モバイルSINETと接続**
 - セキュアなIoT環境の構築が可能
 - ABCI, Wisteria/BDEC-01など既存・将来設置のスパコンと連携

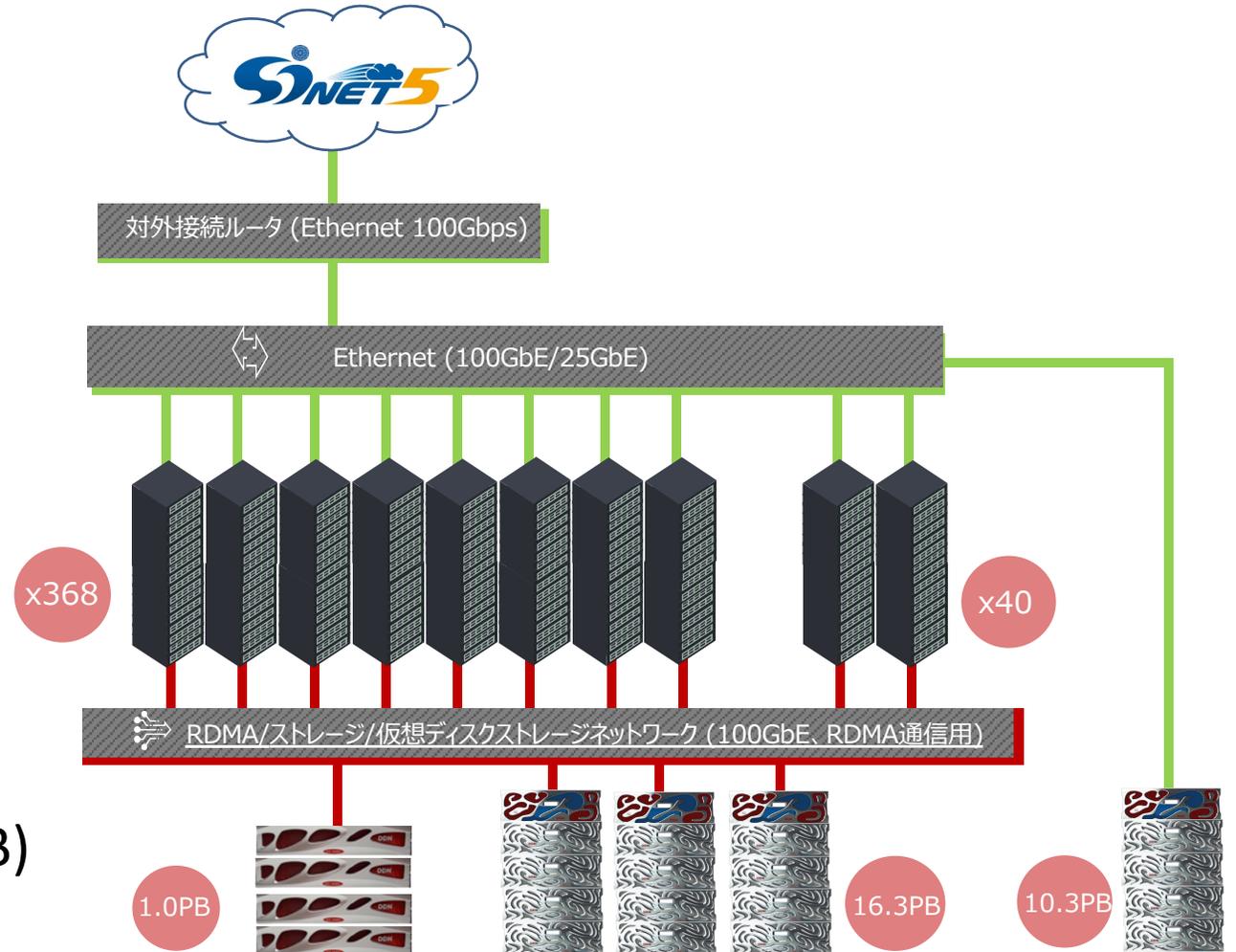
利用イメージ

- ユーザごとに分離された仮想環境を構成
- mdx運用ポータルを通じて仮想環境を操作(VMのデプロイ、設定、etc)
 - 運用ポータルへのログインはNIIの運用する学認と連携(予定)



物理構成

- 2つのネットワーク
 - 外部接続ネットワーク
 - SINETと100G x2で接続
 - 内部高速ネットワーク
 - RDMA
- 複数のストレージ
 - 高速ストレージ (1PB, NVMe)
 - 内部ストレージ (16PB)
 - オブジェクトストレージ(10PB)



計算ノード構成

- 汎用CPUノード, 368台
 - Intel Xeon Platinum 8368 (IceLake-SP) x2ソケット
- GPUノード, 40台
 - Intel Xeon Platinum 8368 (IceLake-SP) x2ソケット + Nvidia A100 x8 GPU
- Hyper Visor
 - VMware vSphere



汎用CPUノード
富士通CX2550 M6



GPUノード
富士通GX2570 M6

- 外部接続ネットワーク向けNICは25Gbps
- 内部高速ネットワーク向けNICは100Gbps

2種類のネットワーク

- 外部接続ネットワーク

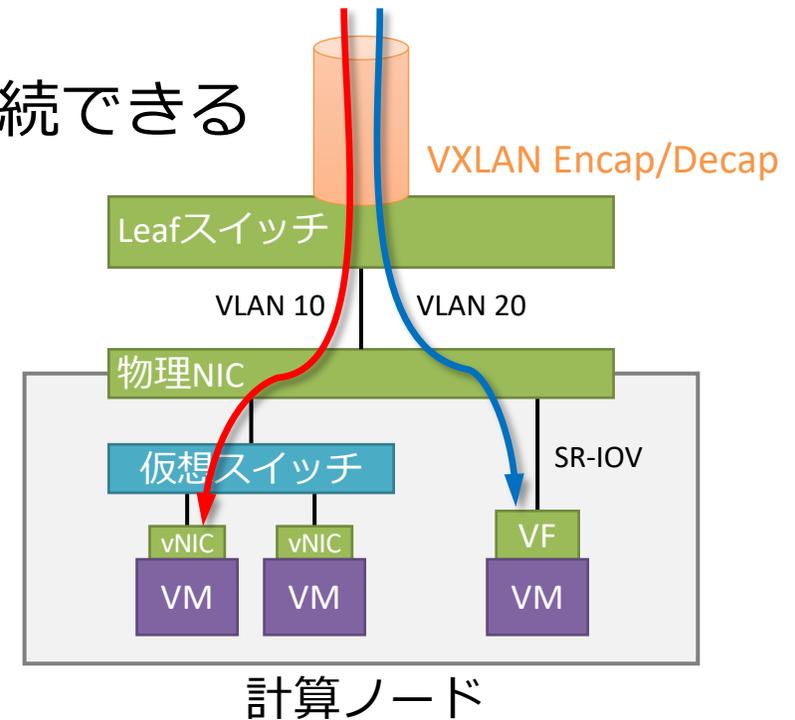
- VMがEthernet/IPで外部と通信するためのネットワーク
 - Spine/LeafトポロジでVXLANオーバーレイ
- 構成機器
 - 外部接続ルータ: MX480 (MPC10E) x2, 400G ready!
 - Spineスイッチ: QFX10002-72Q x4, 100G
 - Leafスイッチ: QFX5120, Downlink 25G, Uplink 100G x4
 - CGN: SRX4600

- 内部高速ネットワーク

- VMがストレージやVM間でRDMAで通信するためのネットワーク
 - Spine/LeafトポロジでVXLANオーバーレイ
- 構成機器
 - Spineスイッチ: SN3700 x16, 200G
 - Leafスイッチ: SN3700, Downlink 100G, Uplink 200G x16

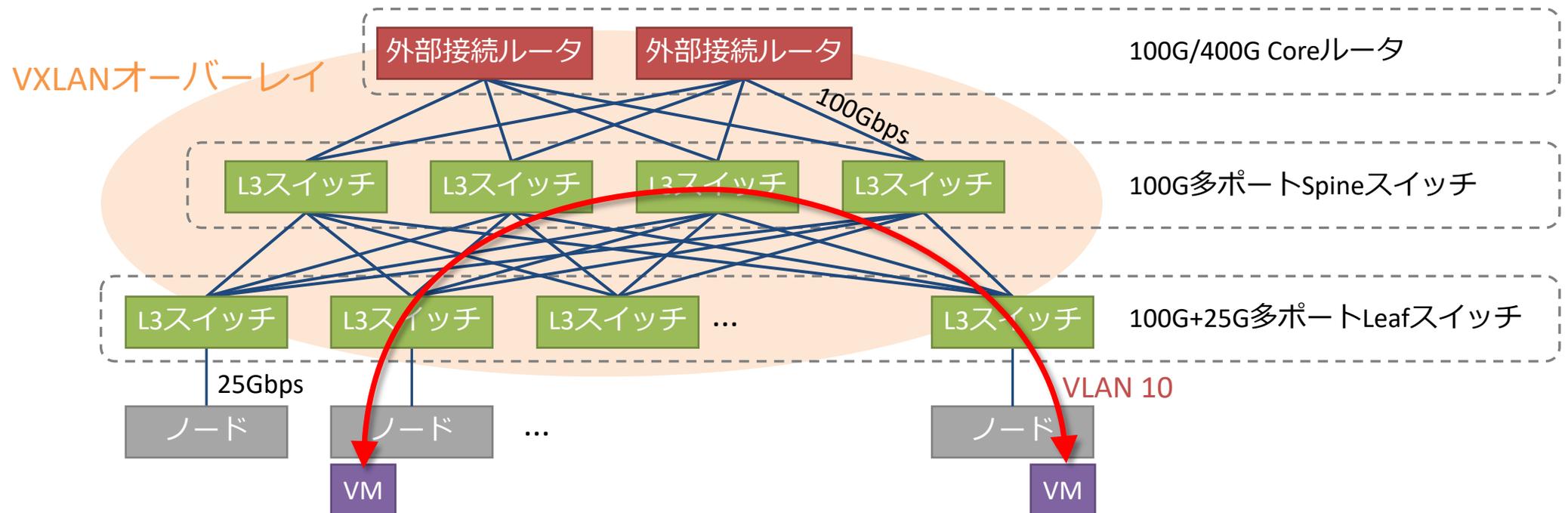
オーバーレイを使う利点

- IPネットワークの利点を維持したままVLANを延伸
 - 複数パスを使ったロードバランスと冗長
 - ユーザのVMがどのノードにおいても同じVLANに接続できる
- 仮想マシンのネットワーク性能と分離
 - トンネルやルーティングの処理をHVのS/Wではなく外部のH/Wスイッチで実施
 - 仮想マシンのNICが仮想NICでもSR-IOVでもネットワーク側は1つのロジックで制御できる
 - 万が一HVが乗っ取られても、他のVLANにはアクセスできない



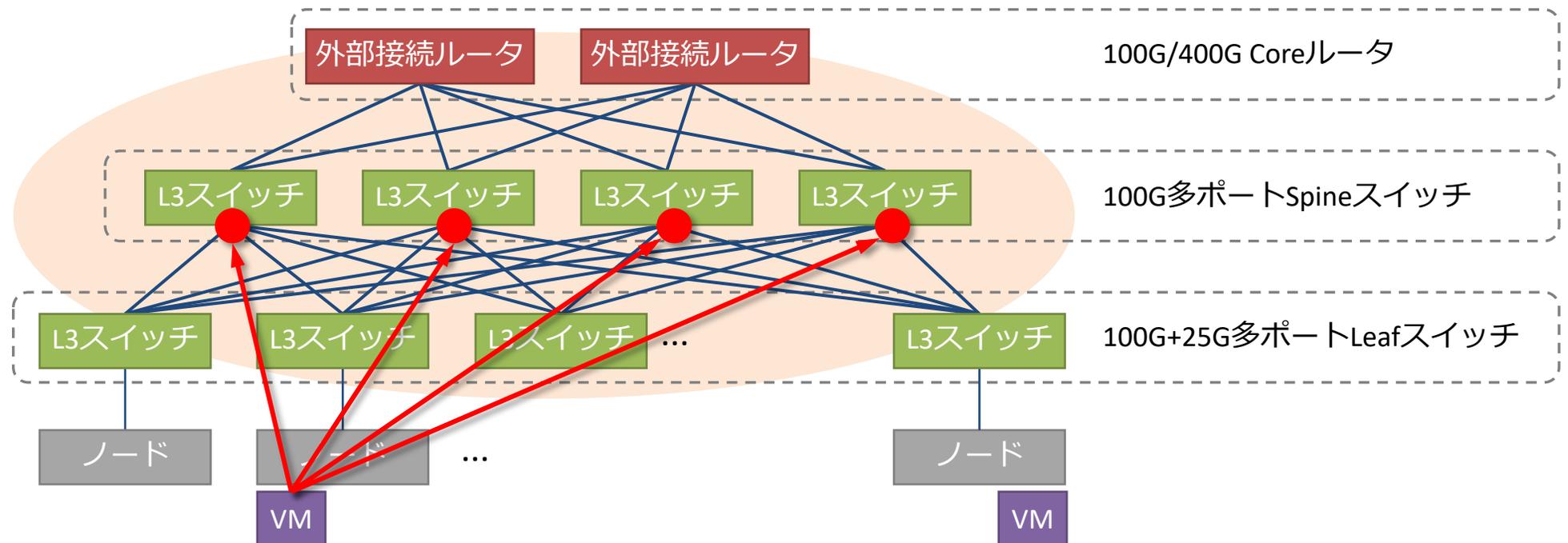
外部接続ネットワーク

- Virtual eXtensible LAN (VXLAN)によるオーバーレイネットワーク
 - データセンター向けの仮想ネットワーク技術
 - 物理はSpine/LeafのIP Clos Fabric ネットワーク
 - コントロールプレーンはEthernet VPN (RFC8365)、 UnderlayはOSPFv2 + BFD



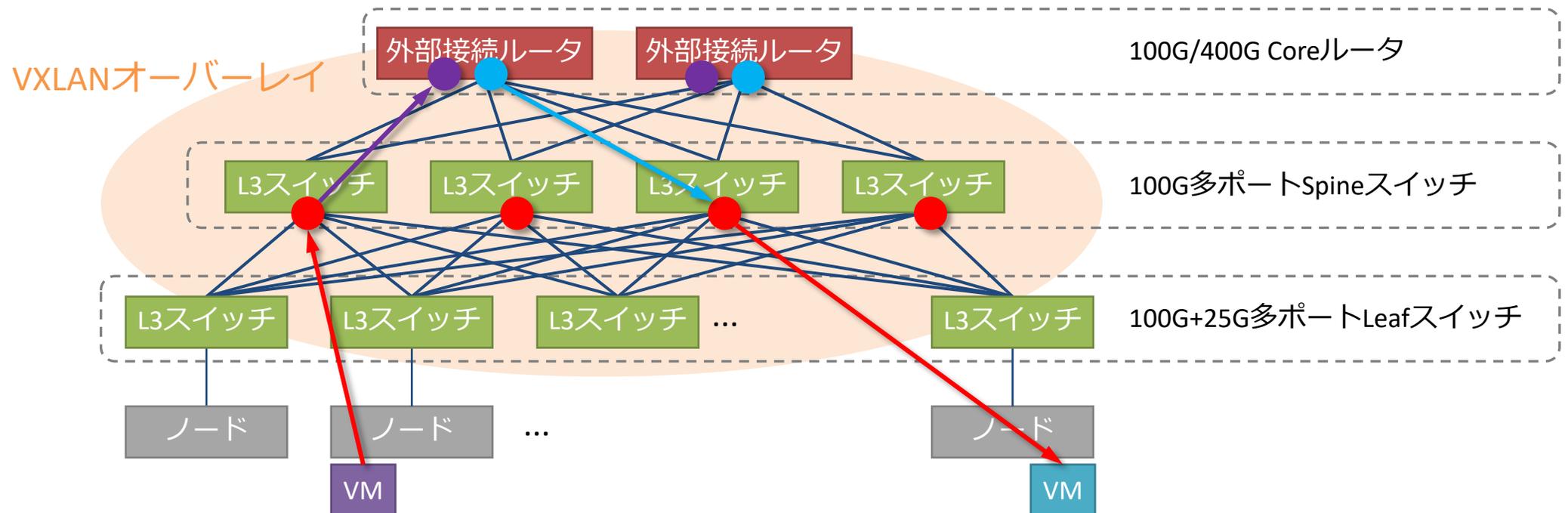
外部接続ネットワーク: ルーティング編

- VM収容セグメントのDefault Gateway
 - VXLAN Anycast Gateway
 - Spineスイッチ4台で冗長+負荷分散



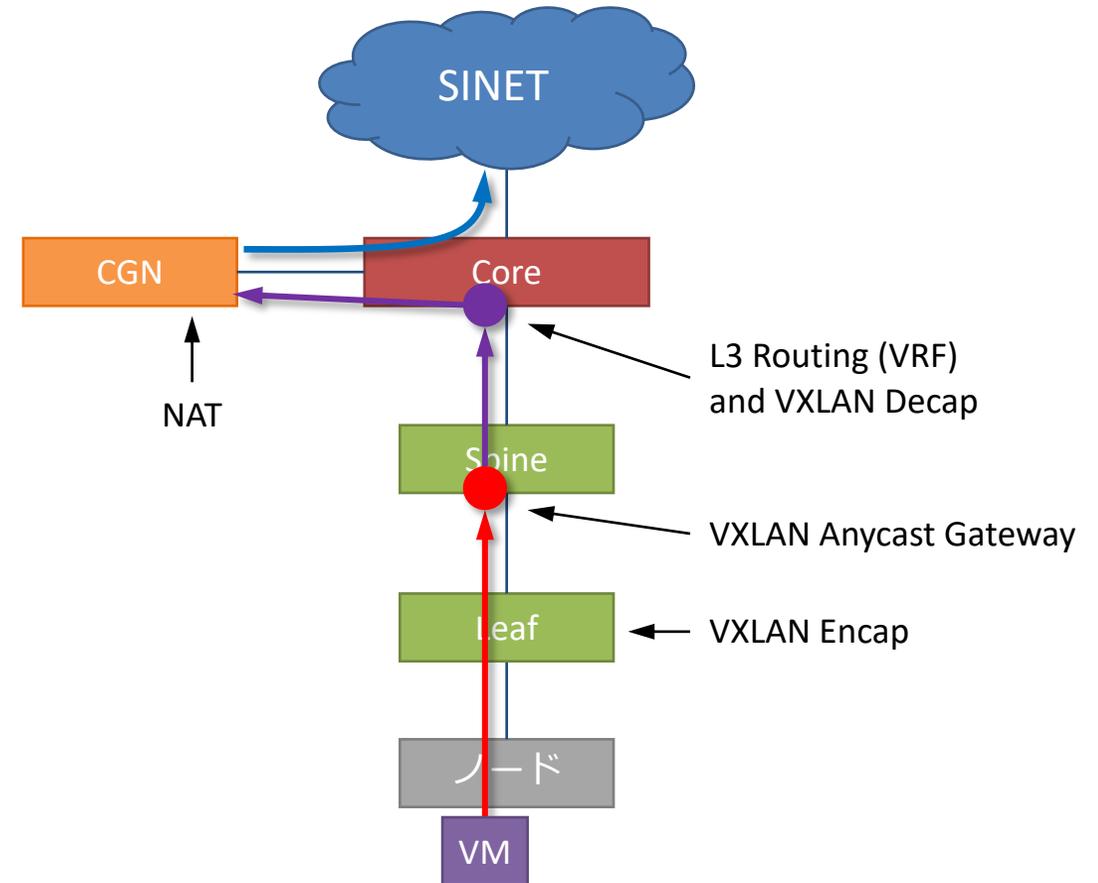
外部接続ネットワーク: ルーティング編

- VM収容セグメント間、及び外部とのルーティング
 - Centrally Routed Bridging (CRB): mdxではCoreでVNI間ルーティングを実施
 - ACLを容量の大きいCoreルータに集約



外部接続ネットワーク: ルーティング編

- Carrier Grade NAT
 - VMにアサインされるのはIPv4 Private、IPv6 Global Unicast
 - コアルータに接続するSRXでCGN
 - VM -> the Internet: Source NAT
 - The Internet -> VM: Destination NAT
 - Core/CGN間はper-VRF BGP



ユーザはポータル画面から設定を変更

- ACLの設定やグローバルv4アドレスの割り当て

upa-test-normal-4

IPv4 IPv6

+ RECORD DELETE EDIT

	Srcアドレス	SrcPrefix長	Srcポート	Dstアドレス	DstPrefix長	Dstポート	プロトコル
<input checked="" type="radio"/>	0.0.0.0	0		10.13.200.0	21		ICMP
<input type="radio"/>	130.69.0.0	16	Any	10.13.200.0	21	Any	TCP

更新

Records per page 10 1 - 2 of 2 Records

+ DNAT DELETE EDIT

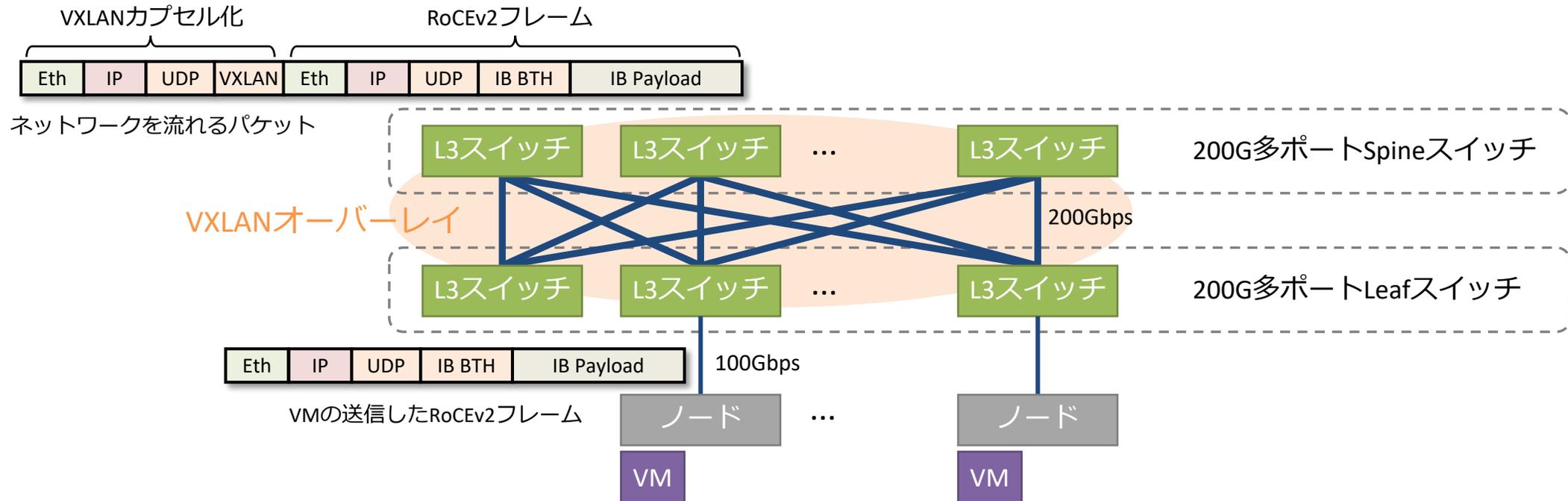
	転送元グローバルIPv4アドレス	セグメント	転送先プライベートIPアドレス
<input checked="" type="radio"/>	163.220.176.215	upa-test-normal-4	10.13.200.118

更新

Records per page 10 1 - 1 of 1 Records

内部高速ネットワーク

- RDMA over Converged Ethernet (RoCE) over VXLAN
 - RoCE: RDMAをIPネットワーク越しに行う技術
 - 外部接続ネットワークと同様、IP ClosトポロジにVXLANを利用
 - Infinibandでは十分なテナント間の分離を実現できなかった
 - コントロールプレーンはEthernet VPN、UnderlayはBGP unnumbered



異なるノード上のVM間通信

```
upa@ubuntu2:~$ iperf3 -c 10.5.236.26
Connecting to host 10.5.236.26, port 5201
[ 5] local 10.5.232.103 port 37812 connected to 10.5.236.26 port 5201
[ ID] Interval          Transfer    Bitrate    Retr  Cwnd
[ 5]  0.00-1.00    sec  2.51 GBytes  21.6 Gbits/sec  414  1.89 MBytes
[ 5]  1.00-2.00    sec  2.72 GBytes  23.3 Gbits/sec   0  1.94 MBytes
[ 5]  2.00-3.00    sec  2.73 GBytes  23.5 Gbits/sec  28  1.76 MBytes
[ 5]  3.00-4.00    sec  2.73 GBytes  23.4 Gbits/sec   0  1.82 MBytes
[ 5]  4.00-5.00    sec  2.73 GBytes  23.5 Gbits/sec   0  1.85 MBytes
```

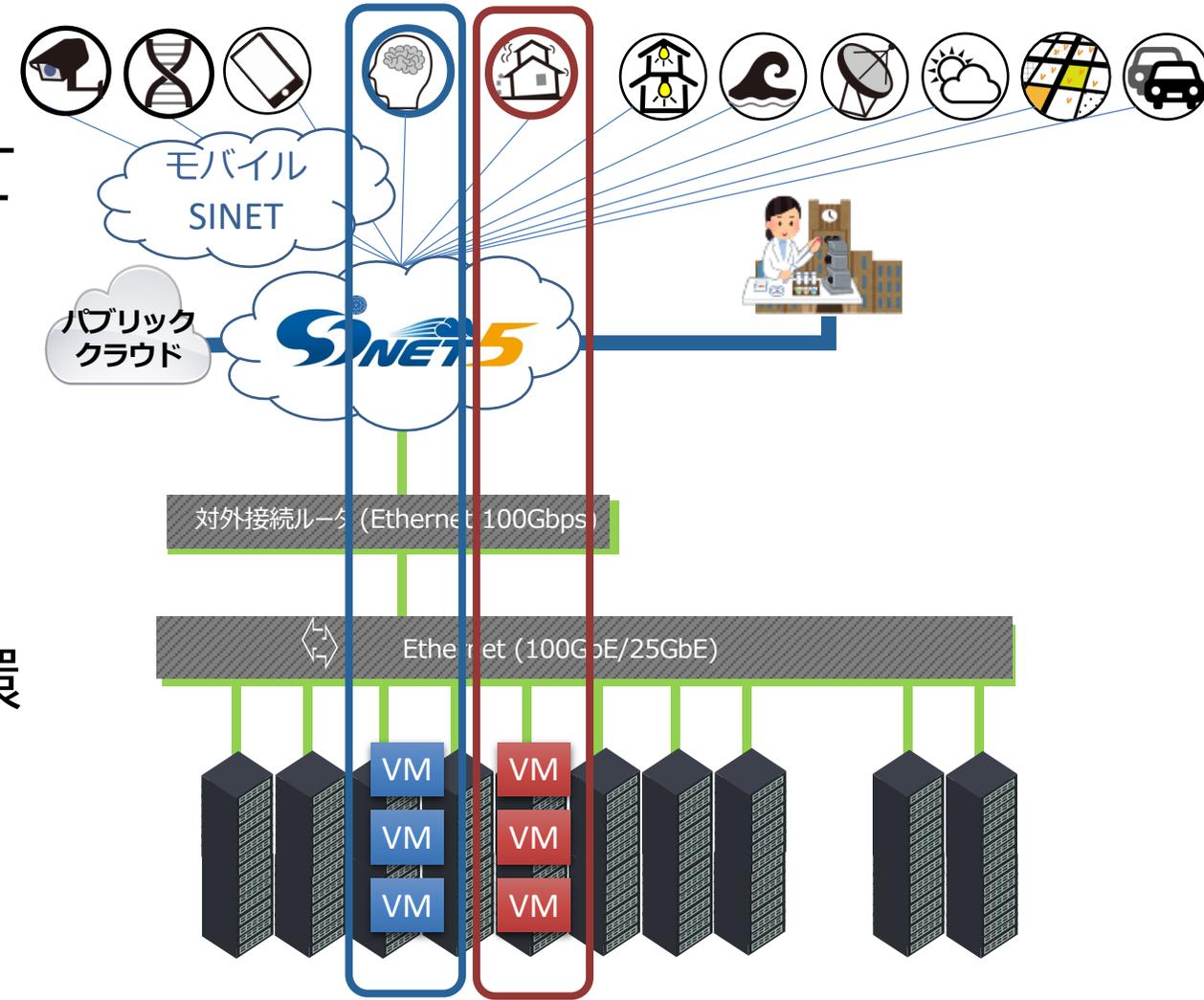
- 外部接続ネットワーク (25Gbps NIC)
 - 仮想NIC (vmxnet3)
 - iperf3で 23.5 Gbps

```
-----
local address: LID 0000 QPN 0x0d0a PSN 0xf3069a
GID: 00:00:00:00:00:00:00:00:00:00:255:255:10:133:236:21
remote address: LID 0000 QPN 0x050a PSN 0xc59591
GID: 00:00:00:00:00:00:00:00:00:00:255:255:10:133:236:20
-----
#bytes    #iterations    BW peak [Gb/sec]    BW average [Gb/sec]    MsgRate [Mpps]
65536     1000            88.89                88.87                    0.169497
```

- 内部高速ネットワーク (100Gbps NIC)
 - SR-IOV (mlx5)
 - ib_send_bwで 88.87 Gbps

SINETとの連携

- SINETのL2VPNサービスを通じてセキュアな閉域網経由でVLANをmdxにもちこむことが可能
- 他のSINETサイト(大学・研究機関)とセキュアに接続
- モバイルSINET経由で、IoTデバイスからmdx上の計算環境へセキュアにデータを転送



まとめと、利用について

- NII, AISTと9大学(北大、東北大、筑波大、東大、東工大、名大、京大、阪大、九大)は共同で、
 - データ科学・活用のための基盤 **mdx** を導入
 - 共同研究・産官学連携の仕組みを運用
- mdx
 - 幅広い用途に使えるIaaS型の仮想化基盤
 - 計算資源、高速ネットワーク、大容量ストレージの提供
 - SINETとの連携による外部ネットワーク接続
- 利用について => <https://mdx.jp>
 - 現在テスト利用が始まっています
 - データ活用、データ提供、プラットフォーム構築、高性能計算・データ処理など、幅広い用途、分野のみなさまに利用できるように